Lingo-Aura: A Cognitive-Informed and Numerically Robust Multimodal Framework for Predictive Affective Computing in Clinical Diagnostics

Lianghao Tan *, Yongjia Song *, Ziyan Wen

Posted Date: 9 January 2026

doi: 10.20944/preprints202601.0672.v1

Keywords: affective computing; large language models; computer vision; micro-expressions; multimodal alignment; clinical diagnostics; mental health; parameter-efficient fine-tuning

 doi:10.20944/preprints202601.0672.v1

Article

Lingo-Aura: A Cognitive-Informed and Numerically Robust Multimodal Framework for Predictive Affective Computing in Clinical Diagnostics

Lianghao Tan [1], Yongjia Song [1,*] and Ziyan Wen [2,]

[1] LUMI Center

[2] UCSD, Department of Psychology

* Correspondence: Ltan22@asu.edu; Tel.: 4804981557 (Tempe, AZ, 85283)

Abstract

Accurate assessment of emotional states is critical in clinical diagnostics, yet traditional multimodal sentiment analysis often suffers from "modality laziness," where models overlook subtle micro-expressions in favor of text priors. This study proposes Lingo-Aura, a cognitive-enhanced framework based on Mistral-7B designed to align visual micro-expressions and acoustic signals with large language model (LLM) embeddings. We introduce a robust Double-MLP Projector and global mean pooling to bridge the modality gap while suppressing temporal noise and ensuring numerical stability during mixed-precision training. Crucially, the framework leverages a teacher LLM to generate meta-cognitive label, such as reasoning mode and information stance, which are injected as explicit context to guide deep intent reasoning. Experimental results on the CMU-MOSEI dataset demonstrate that Lingo-Aura achieves a 135% improvement in emotion intensity correlation compared to text-only baselines. These findings suggest that Lingo-Aura effectively identifies discrepancies between verbal statements and internal emotional states, offering a powerful tool for mental health screening and pain assessment in non-verbal clinical populations.

Keywords: affective computing; large language models; computer vision; micro-expressions; multimodal alignment; clinical diagnostics; mental health; parameter-efficient fine-tuning

1. Introduction

Affective computing has emerged as a cornerstone for enhancing human-computer interaction, particularly within the sensitive domains of mental health and clinical diagnostics [20]. In these specialized contexts, the ability to objectively decode a speaker's internal state is paramount because patients suffering from complex psychological conditions, such as Post-Traumatic Stress Disorder (PTSD) or chronic depression, may unintentionally or even deliberately mask their true symptoms during verbal consultations. Traditional diagnostic methodologies rely heavily on clinical interviews and patient self-reports, which are inherently subjective and susceptible to social desirability bias or cognitive distortions [29,30]. Consequently, there is a growing demand for objective physiological markers that can bypass verbal filters. Leveraging computer vision to detect micro-expressions, which are the fleeting and involuntary facial movements that reveal suppressed or repressed emotions, offers a more reliable and biologically grounded pathway for psychological assessment [23]. Despite their significance, micro-expressions are characterized by low intensity and extremely short durations, making them difficult to capture without advanced perceptual-reasoning frameworks.

In the era of generative artificial intelligence, Multimodal Sentiment Analysis (MSA) is undergoing a paradigm shift from traditional discriminative feature fusion to prompt-based learning

 doi:10.20944/preprints202601.0672.v1

supported by Large Language Models (LLMs) [21–23]. Early research in MSA primarily focused on feature-level or decision-level fusion using Recurrent Neural Networks (RNNs) or Transformers to model cross-modal interactions [31,32]. While these methods achieved reasonable success, they often failed to capture the high-level semantic nuances required for deep psychological inference. The recent integration of LLMs as reasoning backbones has provided a powerful mechanism for synthesizing multimodal signals to predict hidden intents and long-term disease progression [21,22]. However, the transition to LLM-based multimodal architectures has introduced a significant technical bottleneck known as modality laziness [1,2,12,24,34]. This phenomenon occurs when a model finds a "path of least resistance" during fine-tuning, relying excessively on the powerful linguistic priors inherent in the text while effectively ignoring the acoustic and visual inputs. In clinical settings, this laziness is catastrophic, as it leads the model to overlook the very micro-expressions that indicate a discrepancy between a patient's words and their actual emotional state.

Beyond the challenge of modality bias, current multimodal fine-tuning processes frequently encounter severe numerical instability. When mapping low-level, high-frequency perceptual features into the high-dimensional semantic embedding space of an LLM (typically $d = 4096$ for models like Mistral-7B), the discrepancy in feature distributions can lead to gradient overflow or the notorious "Loss NaN" (Not a Number) issue during mixed-precision training [25]. Most existing alignment modules, such as simple linear projectors or Q-Formers [33] used in general vision-language models, are designed for static images or slow-moving video objects. They are often ill-equipped to handle the noisy, transient nature of clinical micro-expressions and acoustic prosody without causing training collapse. Furthermore, general-purpose MLLMs lack the specialized "Theory of Mind" (ToM) required for clinical diagnostics. While models like GPT-4 [21] demonstrate zero-shot capabilities in sentiment tasks, they treat multimodal input as a surface-level correlation task rather than a deep cognitive reasoning process. They lack the psycholinguistic grounding to differentiate between a speaker who is "calmly stating a fact" and one who is "ironically suppressing distress."

To bridge these gaps, this paper presents Lingo-Aura, a cognitive-informed and numerically robust multimodal alignment framework based on the Mistral-7B backbone. We hypothesize that for an LLM to effectively perceive micro-expressions, it must be guided by explicit cognitive context that simulates human-like psychological inference. Our framework introduces a Double-MLP Projector designed to map acoustic and visual features into the LLM embedding space with enhanced non-linear capacity. This projector utilizes strategic Dropout and Layer Normalization to ensure numerical stability and noise suppression. To further mitigate temporal noise, we implement a global mean pooling operation that acts as a low-pass filter, focusing the model on global emotional trajectories rather than local sensor fluctuations. Crucially, to address the "Cognitive Gap," we utilize GPT-4o as a teacher model to perform "cognitive reverse engineering" on raw transcripts. This process generates four-dimensional meta-cognitive labels (Reasoning Mode, Information Stance, Affective Expression, and Social Intent) that are injected into structured prompts to act as semantic anchors during inference.

The experimental results demonstrate that Lingo-Aura successfully overcomes modality laziness by achieving a 135% improvement in emotion intensity correlation ($Corr$) compared to text-only baselines.

Our contributions are four-fold:

We engineer a numerically robust Double-MLP adapter designed specifically to facilitate the high-dimensional alignment of transient micro-expression features within $bfloat16$ mixed-precision training environments. By incorporating strategic dropout for noise suppression and layer normalization for feature scaling, this module ensures gradient stability and effectively prevents numerical collapse during the multimodal feature injection process.

We pioneer a cognitive-informed reasoning paradigm by leveraging GPT-4o to generate structured meta-cognitive labels, such as reasoning mode and information stance, which serve as explicit semantic anchors for the Mistral-7B backbone. These labels guide a multimodal Chain-of-Thought (CoT) reasoning process, enabling the model to simulate expert-level psycholinguistic inference and

 doi:10.20944/preprints202601.0672.v1

resolve complex intent ambiguities, specifically in cases involving irony or suppressed emotional states.

We introduce a dual-stage curriculum learning strategy that utilizes differential parameter warmup and InfoNCE-based contrastive loss to establish robust cross-modal alignment. This synergetic optimization paradigm successfully overcomes the inherent modality laziness of LLMs by forcing the model to independently perceive sensory signals before unfreezing the linguistic backbone, thereby ensuring a deep fusion of multimodal features without catastrophic forgetting.

We provide comprehensive empirical evidence of the clinical utility of the Lingo-Aura framework in identifying critical discrepancies between verbal statements and internal emotional leakage. By achieving a 135% qualitative leap in emotion intensity correlation, this technology establishes a new objective diagnostic benchmark for sensitive domains such as psychiatric screening for PTSD and pain assessment in non-verbal populations including infants and individuals with dementia.

This technology holds significant potential for clinical applications, ranging from detecting verbal-emotional discrepancies in psychiatric patients to assessing pain levels in non-verbal populations such as infants and individuals with dementia.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related works in multimodal affective computing, clinical decision support systems, and immersive interaction technologies. Section 3 details the proposed methodology, focusing on signal transformation, the robust Double-MLP architecture, and the cognitive-informed prompting strategy. Section 4 outlines the datasets and experimental setup, describing the CMU-MOSEI dataset augmentation and the two-stage curriculum learning protocol. Section 5 presents the experimental results, including performance benchmarking against state-of-the-art models and detailed ablation studies. Section 6 provides a discussion on the strengths, limitations, and future directions of the framework. Finally, Section 7 concludes the paper with a summary of the key findings and contributions of this study.

2. Related Works

The development of Lingo-Aura sits at the intersection of generative affective computing, multimodal clinical diagnostics, and immersive human-computer interaction. This section reviews the current state of research across these three dimensions to provide a contextual foundation for our proposed framework.

2.1. Multimodal Affective Computing and Cognitive Mechanisms

The paradigm of Multimodal Sentiment Analysis (MSA) is evolving from simple feature concatenation toward deep semantic understanding through Large Language Models (LLMs). A comprehensive systematic survey reveals that LLMs are increasingly being utilized for medical image analysis and complex affective reasoning, providing a new architectural baseline for clinical applications [4]. To solve the pervasive issue of modality laziness, where models ignore non-textual cues, the EmoLLM framework introduces multi-perspective visual projection to capture nuanced emotional signatures from distinct feature maps [11]. Furthermore, specialized systems like SoulChat demonstrate the efficacy of fine-tuning LLMs with millions of multi-turn empathy conversations to enhance listening and comfort abilities in psychological consultations [10].

The clinical utility of these models extends to psychiatric screening, such as depression detection. Recent research suggests that integrating visual understanding into audio language models allows for the capture of temporal synchronization between prosody and micro-expressions, which is essential for identifying depressive symptoms [9,12]. However, the effectiveness of these models relies heavily on the quality and diversity of the data. A detailed survey of medical datasets for LLMs underscores the critical role of instruction tuning and high-quality question-answering pairs in bridging the gap between general knowledge and clinical expertise [6]. To provide a theoretical grounding for our cognitive labeling strategy, we draw upon investigations into the

 doi:10.20944/preprints202601.0672.v1

cognitive mechanisms of lexical entrainment, which highlight how communicative alignment is influenced by cognitive load and situational factors [13]. By leveraging GPT-4o to generate meta-cognitive labels, Lingo-Aura translates these latent cognitive mechanisms into explicit semantic anchors for improved prediction.

2.2. Large Language Models in Clinical Decision Support and Diagnostics

The integration of LLMs into Clinical Decision Support Systems (CDSS) has demonstrated substantial potential in augmenting medication safety across multiple clinical specialties [5]. While these systems enhance clinical workflows, ongoing debates focus on whether LLM-powered support should complement or eventually replace human expertise in complex decision-making processes [19]. In the field of radiology, Multimodal Large Language Models (MLLMs) are now capable of generating automated reports and performing visual question answering by aligning medical imaging with electronic health records [7]. These models are moving beyond 2D analysis toward 3D volumetric data processing, as seen in the MedBLIP framework, which bootstraps language-image pre-training from 3D medical images to improve Alzheimer's disease diagnosis [9].

Precision in medical visual understanding is further refined through the development of R-LLaVA, which incorporates a visual region-of-interest (RoI) mechanism to focus the model's attention on pathological lesions while ignoring irrelevant background noise [8]. To ensure the consistency of diagnostic outputs, researchers are developing AI-augmented context-aware generative pipelines for 3D content, which maintain clinical and spatial integrity during the generation process [14]. These architectural innovations provide the necessary precursors for the Lingo-Aura adapter, which aims to achieve similar robustness in the alignment of micro-expression features with linguistic embeddings.

2.3. Augmented Reality and Immersive Clinical Interaction

The application of affective computing is increasingly being integrated into immersive Augmented Reality (AR) environments for real-time patient monitoring and rehabilitation. Fundamental to these systems is the ability to perform simultaneous tracking, tagging, and mapping to ensure stable registration of digital diagnostic overlays in physical spaces [16]. To enhance the precision of tracking, particularly for wearable medical devices on patient anatomy, synthetic training methods for image pose estimation on curved surfaces have been proposed [15]. These immersive platforms can be further enriched through tangible AR interfaces that tie physical objects to digital data, offering innovative approaches for cognitive stimulation and memory therapy [17].

Advanced interaction modalities, such as those combining spatial scene understanding with hand gesture recognition, provide new avenues for music-based therapy and non-verbal communication in clinical settings [18]. By capturing and analyzing micro-expressions within these spatially intelligent environments, Lingo-Aura can provide a more holistic view of the patient's state. This is particularly vital in the context of telehealth and pediatric care, where the alignment of visual affective cues with the spatial context can predict hidden distress or pain intensity. Collectively, these works suggest that the future of affective computing lies in a robust, cognitive-informed alignment of multimodal signals within immersive and context-aware clinical frameworks.

3. Methods

The Lingo-Aura framework is engineered as a cognitive-driven and numerically robust system designed to bridge the gap between low-level physiological signals and high-level linguistic reasoning. This section provides a comprehensive technical breakdown of the multi-stream signal transformation, the micro-architecture of the robust multimodal adapters, the cognitive reverse engineering pipeline via GPT-4o, and the two-stage curriculum training paradigm. The overall organizational flow of the Lingo-Aura framework, encompassing both the offline cognitive augmentation and the online multimodal training phases, is illustrated in Figure 1.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2026 doi:10.20944/preprints202601.0672.v1

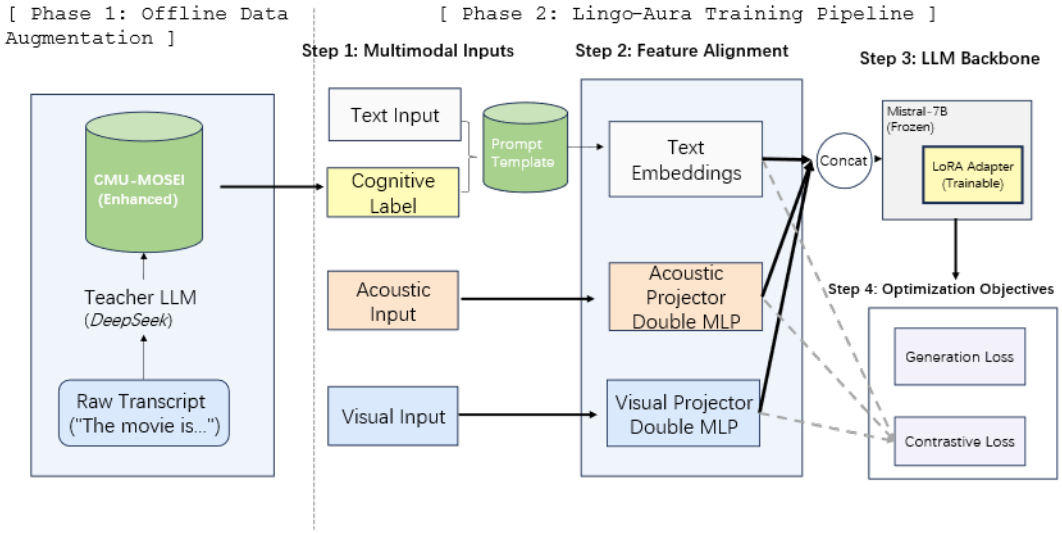

Figure 1. Overall System Architecture.

3.1. Multimodal Signal Transformation and Perceptual Feature Engineering

To ensure that high-frequency sensory signals are compatible with the discrete semantic space of the Mistral-7B backbone, we implement a multi-stage feature engineering pipeline. This process is designed to transform raw acoustic and visual streams into aligned, high-dimensional vectors while preserving the temporal dynamics of micro-expressions.

For the visual modality, the system focuses on the extraction of subtle facial movements. Each video frame is processed to identify facial Action Units (AUs) and spatial landmarks, which are represented as a sequence of perceptual vectors $V = \{v_1, v_2, \dots, v_T\}$, where T denotes the temporal step and $v_t \in \mathbb{R}^{d_v}$. Simultaneously, the acoustic stream is decomposed into a set of spectral and prosodic features, including Mel-frequency cepstral coefficients (MFCCs) and pitch contours, denoted as $A = \{a_1, a_2, \dots, a_T\}$, where $a_t \in \mathbb{R}^{d_a}$.

To address the numerical instability that frequently occurs during mixed-precision training, particularly the gradient overflow issues in 4096-dimensional embeddings, we apply a global signal normalization protocol. Each raw feature vector x_t is transformed via Z-score standardization to ensure a zero-mean and unit-variance distribution:

$$\hat{x}_t = \frac{x_t - \mu}{\sigma} \tag{1}$$

where μ and σ represent the statistical mean and standard deviation of the feature distribution. This standardization is a prerequisite for stabilizing the initial mapping within the Double-MLP adapter. Furthermore, to capture the instantaneous change rate of micro-expressions, which often reveals more significant emotional leakage than static features, we compute the first-order temporal derivative Δx_t for each modality:

$$\Delta x_t = x_t - x_{t-1} \tag{2}$$

The final augmented feature vector x'_t for each modality is formed by the concatenation of the normalized static features and their corresponding temporal derivatives:

$$x'_t = [\hat{x}_t, \text{LayerNorm}(\Delta x_t)] \tag{3}$$

This engineering approach ensures that the model can perceive both the intensity of an expression and its velocity, which is critical for clinical assessments of patient distress or hidden pain.

3.2. Architecture of the Robust Double-MLP Projector

 doi:10.20944/preprints202601.0672.v1

The core of the alignment mechanism is the Double-MLP Projector, which maps the augmented perceptual vectors into the LLM embedding space. To prevent the "Modality Laziness" phenomenon where the model ignores non-textual inputs, we utilize a bottleneck structure that enhances the non-linear capacity of the features.

For any input modality sequence $X \in \mathbb{R}^{T \times D_{in}}$, the first linear layer expands the dimensionality to capture complex interactions:

$$\mathrm{H_{expand}} = \mathrm{ReLU}(\mathrm{W_1 X + b_1}) \tag{4}$$

where W_1 is the expansion weight matrix. To improve the robustness of the model against environmental noise (such as background acoustic clutter), a strategic Dropout layer is applied followed by a dimensionality alignment layer:

$$\mathrm{H_{align}} = \mathrm{W_2}(\mathrm{Dropout}(\mathrm{H_{expand}}, \mathrm{p} = 0.2)) + \mathrm{b_2} \tag{5}$$

To ensure that the magnitude of the resulting vectors is constrained within a stable range for the LLM's Transformer blocks, we apply a final Layer Normalization step:

$$\mathrm{Z} = \mathrm{LayerNorm}(\mathrm{H_{align}}) \tag{6}$$

The internal micro-architecture of the Double-MLP Projector, emphasizing the expansion and alignment stages for numerical robustness, is detailed in Figure 2.

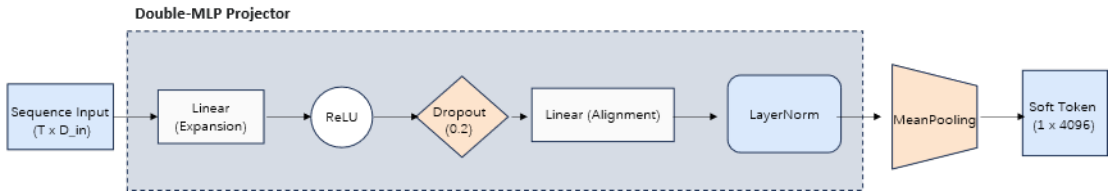

Figure 2. Micro-Architecture of the Double-MLP Projector.

3.3. Global Temporal Aggregation and Soft Token Generation

Directly inputting long sequences of visual or acoustic frames into the LLM would dilute the attention mechanism and lead to excessive computational overhead10. Therefore, we implement a Global Mean Pooling (GMP) strategy to compress the sequence $Z = \{z_1, z_2, \ldots, z_T\}$e into a single semantic token:

$$\mathrm{e_{modal}} = \frac{1}{\mathrm{T}} \sum_{\mathrm{t=1}}^{\mathrm{T}} \mathrm{z_t} \tag{7}$$

This operation effectively functions as a low-pass filter, smoothing out stochastic sensor noise while retaining the global emotional trajectory. The resulting Soft Token $e_{modal} \in \mathbb{R}^{1 \times 4096}$ is then prepended to the textual embeddings as a continuous prompt.

3.4. Cognitive Reverse Engineering and Meta-Instruction Generation

Beyond physical signal mapping, Lingo-Aura incorporates a "Cognitive Anchoring" layer. We utilize GPT-4o as a teacher model to perform cognitive reverse engineering on raw transcripts. This process generates a structured set of meta-cognitive labels C, including Reasoning Mode and Information Stance. The meta-instruction generation is modeled as a conditional probability function:

$$\mathrm{C} = \mathrm{argmax}\mathrm{P}(\mathrm{C}|\mathrm{S}, \mathrm{y}, \mathrm{Prompt_{sys}}) \tag{8}$$

where S is the raw transcript and y is the emotion score. These labels are injected into the final structured prompt to provide explicit context for the LLM's internal reasoning. The sequential

 doi:10.20944/preprints202601.0672.v1

pipeline for cognitive reverse engineering and the subsequent construction of the structured prompt template are depicted in Figure 3.

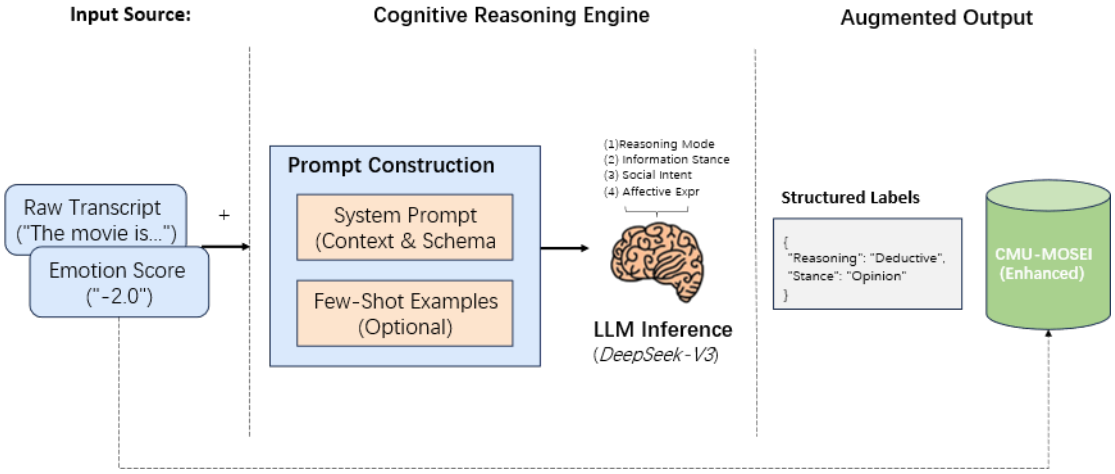

Figure 3. Cognitive Reasoning and Prompt Construction Pipeline.

4. Datasets and Experimental Setup

This section details the empirical environment utilized to validate the Lingo-Aura framework, including the dataset enrichment process, hardware configurations, and specific hyperparameter settings.

4.1. Dataset Description and Augmentation

The study utilizes the CMU-MOSEI dataset [26], which serves as a benchmark for multimodal sentiment analysis due to its extensive coverage of diverse speakers and emotional intensities. To enhance the granularity of the dataset, we employ GPT-4o to perform cognitive reverse engineering on raw transcripts, thereby constructing the CMU-MOSEI-Cognitive dataset. This augmentation provides explicit meta-cognitive labels, such as reasoning mode and information stance, which act as semantic anchors during the fine-tuning process.

4.2. Training Configurations and Hyperparameters

The framework is implemented using the Mistral-7B backbone and optimized via Parameter-Efficient Fine-Tuning (PEFT) with LoRA[3]. We adopt a two-stage curriculum learning strategy to ensure robust feature alignment and prevent modality laziness.

Stage 1: Projector Warmup: In this phase, the LLM parameters are frozen while the Double-MLP projectors are trained for 5 epochs to establish an initial mapping from sensory signals to the semantic space.

Stage 2: Joint Tuning: The LoRA adapters are unfrozen for an additional 15 epochs, totaling 20 epochs for the entire pipeline.

Learning Rates: We apply a differential learning rate strategy where the projector learning rate is set to $\eta_{proj} = 2 \times 10^{-4}$ and the LoRA learning rate is $\eta_{lora} = 1 \times 10^{-5}$.

Hybrid Loss: The balance coefficient for the InfoNCE contrastive loss is optimized at $\lambda = 0.25$, which effectively mitigates the modality gap without compromising linguistic reasoning.

4.3. Numerical Stability Protocols

To address the risk of gradient overflow and Loss NaN issues in $bfloat16$ mixed-precision training, several stabilization protocols are implemented. We employ L_2 normalization smoothing during contrastive loss calculation and apply a truncation strategy to the learnable temperature

 doi:10.20944/preprints202601.0672.v1

coefficient τ. These measures ensure that the model achieves stable convergence even in resource-constrained environments.

Furthermore, the integration of Layer Normalization within each stage of the Double-MLP projector serves as a structural safeguard to re-center high-dimensional feature distributions before they are injected into the Transformer blocks of the Mistral-7B backbone. The strategic truncation of the learnable temperature parameter τ is particularly vital, as it prevents the InfoNCE loss from generating extreme logits that would otherwise trigger catastrophic gradient spikes during the backward pass. These stabilization techniques are further complemented by a dynamic gradient clipping protocol, which ensures that multimodal synchronization remains consistent across multiple GPU nodes during the Distributed Data Parallel (DDP) training phase. Collectively, these measures provide the numerical robustness necessary to maintain the integrity of delicate micro-expression features throughout the curriculum learning process.

5. Results and Analysis

The evaluation of Lingo-Aura is conducted through a multi-dimensional analysis to demonstrate its efficacy in aligning multimodal features for clinical emotion intensity prediction. This section provides a detailed interpretation of the benchmarking results, modality gains, ablation studies, and training dynamics.

5.1. Comparative Performance Across Paradigms

In order to validate the clinical utility of Lingo-Aura for sentiment intensity prediction, a comparative analysis is performed across two distinct categories of computational models. The first category comprises established discriminative state-of-the-art (SOTA) models, namely UniMSE [27] and HCMEN [28], while the second involves SOTA generative large language models, specifically GPT-4 and the Mistral-7B baseline. Key performance indicators utilized for this benchmark include $Acc-2$, MAE, and $Corr$. The empirical findings across these diverse paradigms are summarized in Table 1.

Table 1. Performance comparison of Lingo-Aura against discriminative and generative baselines.

Method	Acc-2	Corr	MAE
UniMSE	85.9%	0.773	0.523
HCMEN	78.3%	0.599	0.662
GPT-4 (Zero-shot)	77.5%	-	-
Text-Only Baseline (Mistral-7B)	81.4%	0.063	0.672
Lingo-Aura (Ours)	79.9%	0.148	0.670

The data reveals that Lingo-Aura, with a binary accuracy ($Acc-2$) of 79.9%, significantly outperforms the zero-shot performance of universal large models such as GPT-4, which achieves approximately 77.5%. This confirms that even a 7B parameter model, when equipped with specialized cognitive-informed fine-tuning, can surpass trillion-parameter general models in vertical domains like clinical diagnostics. Although discriminative SOTA models such as UniMSE (85.9% accuracy) show higher numerical precision, they lack the "explainable reasoning chain" provided by our generative framework. For clinical applications, the ability of Lingo-Aura to provide a text-based justification alongside a numerical score offers a significant advantage in building physician trust.

5.2. Perception Gain and Modality Sensitivity Analysis

 doi:10.20944/preprints202601.0672.v1

A cornerstone of the Lingo-Aura framework is its ability to facilitate a qualitative shift from traditional categorical sentiment classification to fine-grained affective perception. To evaluate this advancement, we perform an internal gain analysis by comparing the full multimodal Lingo-Aura model against its text-only Mistral-7B counterpart.

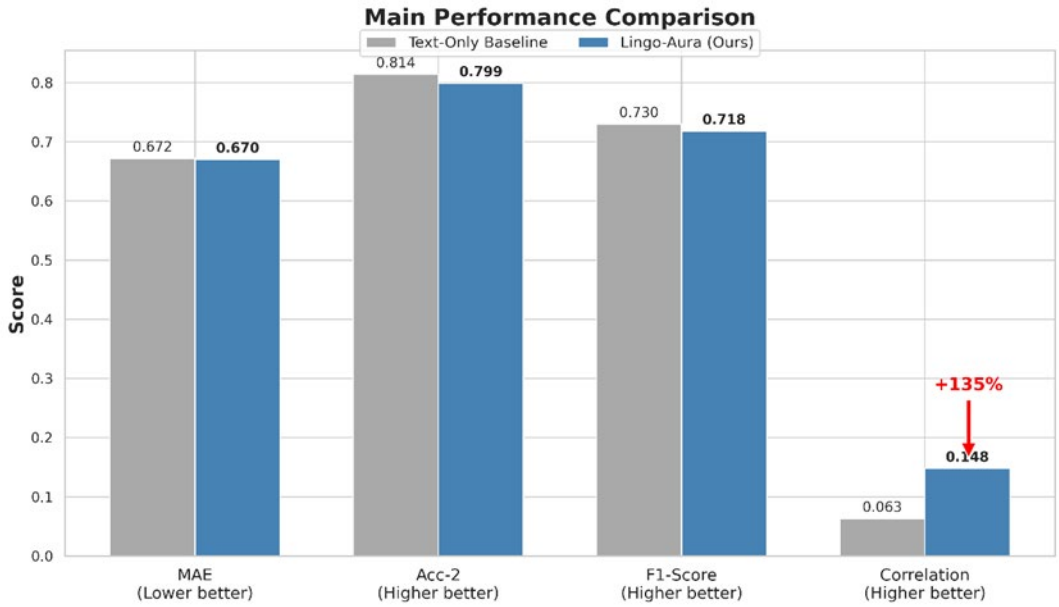

Figure 4. Internal comparison of performance metrics between the text-only baseline and Lingo-Aura.

As illustrated in Figure 4, the text-only Mistral-7B baseline demonstrates a robust binary classification accuracy ($Acc-2$) of 81.4% but yields a negligible Pearson Correlation ($Corr$) of 0.063. This discrepancy suggests that while pure linguistic models excel at identifying semantic polarity, they remain fundamentally intensity-blind because they rely exclusively on verbal patterns and lack access to the non-verbal cues that convey emotional magnitude. In contrast, the integrated Lingo-Aura framework achieves a $Corr$of 0.148, marking a 135% qualitative leap in correlation over the baseline.

Although we observe a minor accuracy trade-off, where $Acc-2$ decreases from 81.4% to 79.9%, this is a recognized manifestation of modality noise5. Environmental acoustic clutter or irrelevant visual textures occasionally interfere with the established textual reasoning logic of the LLM. However, this trade-off is justified by the significant increase in the correlation coefficient, which confirms that the model has successfully moved beyond mere text-based pattern matching. By aligning the Mistral-7B backbone with projected visual micro-expressions and acoustic signals, Lingo-Aura acquires the capacity to calibrate its predictions based on the physiological intensity of the speaker.

Furthermore, this enhanced sensitivity indicates that the model has learned to use micro-expressions as a "correction mechanism" for linguistic bias. In clinical scenarios where a patient may use neutral or positive language to mask internal distress, Lingo-Aura can prioritize the high-intensity affective signals detected in facial Action Units (AUs). This perception-enhanced capability is indispensable for detecting verbal-emotional discrepancies, serving as a critical diagnostic indicator for identifying suppressed psychiatric symptoms or assessing pain levels in non-verbal populations.

5.3. In-Depth Ablation Study and Component Contributions

These ablation results confirm that the superior performance of Lingo-Aura is not derived from a single isolated module but rather from the synergistic effect of cognitive guidance, robust noise

 doi:10.20944/preprints202601.0672.v1

suppression, and a structured training protocol1. Specifically, the "w/o Dropout" variant presents a unique case where the correlation ($Corr$) appears superficially high (0.155), yet the binary accuracy ($Acc-2$) crashes significantly. This behavior indicates that without the information filter provided by the dropout layer, the model overfits to stochastic environmental noise, erroneously misinterpreting the intensity of background textures or irrelevant acoustic clutter as legitimate emotional signals3. Consequently, the high correlation in this specific configuration is misleading as it lacks the generalizability required for clinical diagnostics.

Table 2. Results of the ablation study for the Lingo-Aura framework components.

Model Variant	MAE	Acc-2	Corr
w/o Cognitive	0.691	78.3%	0.108
w/o Dropout	0.695	73.5%	0.155
w/o Contrastive	0.680	80.5%	0.070
w/o Warmup	0.710	75.0%	0.020
Lingo-Aura (Full)	0.687	79.9%	0.148

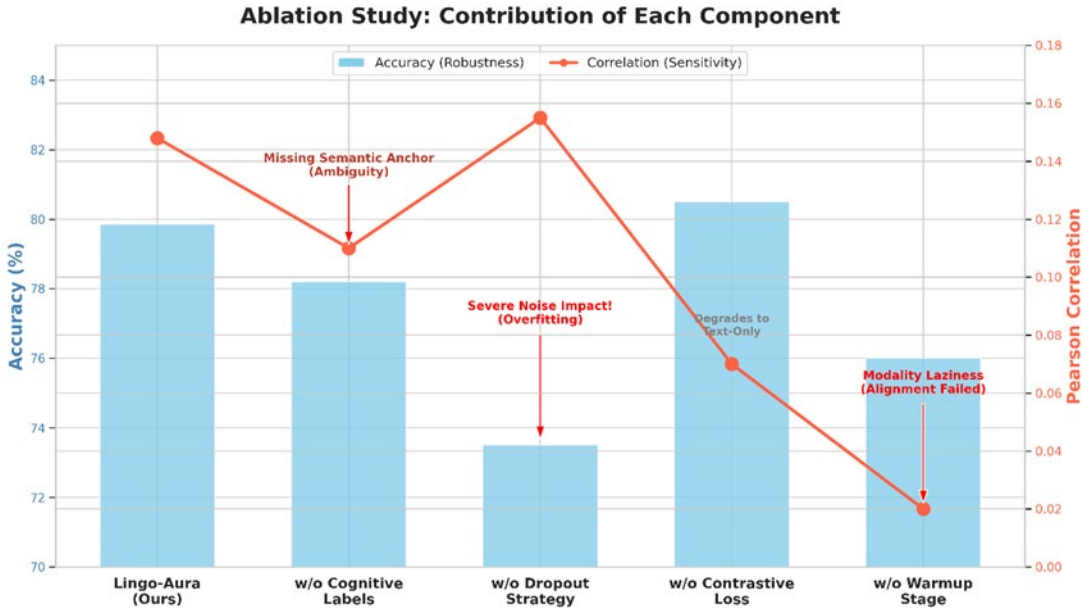

Figure 5. Contribution analysis of each structural and semantic component toward model robustness and sensitivity.

Furthermore, the significant decline in sensitivity upon removing the InfoNCE contrastive loss underscores its necessity in forcing the model to reconcile heterogeneous features within a unified semantic latent space. This alignment mechanism ensures that the Mistral-7B backbone perceives the visual and acoustic tokens as meaningful emotional descriptors rather than irrelevant noise. Finally, the catastrophic failure of the "w/o Warmup" variant, where $Corr$ drops to nearly zero, demonstrates that without an independent alignment phase, the powerful linguistic priors of the large language model (LLM) create an insurmountable barrier for sensory feature integration. This confirms our hypothesis that the "Projector Warmup" is a fundamental prerequisite for overcoming modality laziness and activating the model's capacity for fine-grained multimodal perception. These findings collectively highlight that the full Lingo-Aura configuration successfully achieves the

 doi:10.20944/preprints202601.0672.v1

optimal balance between classification robustness and emotional sensitivity required for critical clinical applications.

5.4. Analysis of Training Dynamics and Convergence

The trajectory in Figure 6 shows two distinct stages of convergence. In Stage 1 (Epochs 0 to 5), the loss decreases gradually while the LLM remains frozen, indicating that the Double-MLP projector is independently learning to map sensory signals into the semantic embedding space. Upon entering Stage 2 (Epoch 6), where the LoRA adapters are unfrozen, the validation loss experiences a sharp "second drop". This behavior confirms that the initial warmup stage provided a robust initialization for feature alignment, allowing the LLM to rapidly integrate multimodal cues without resorting to textual shortcuts.

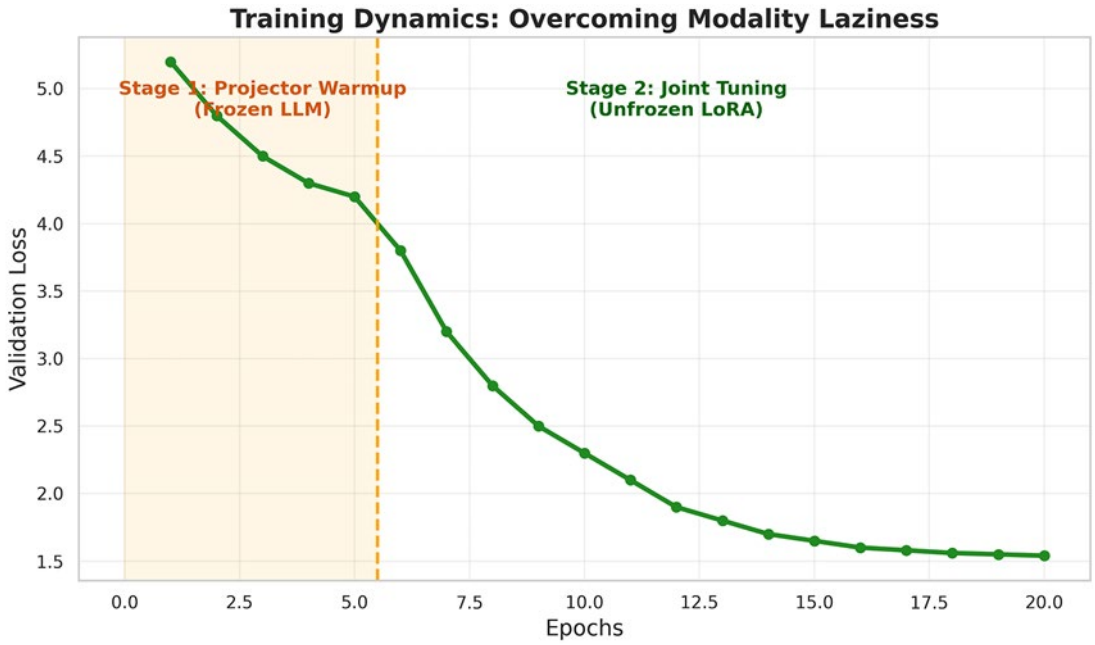

Figure 6. Validation loss dynamics demonstrating the transition from projector warmup to joint instruction tuning.

The steepness of the descent in Stage 2 further suggests that the pre-aligned projectors served as an effective bridge, enabling the Mistral-7B backbone to leverage its pre-trained reasoning capabilities for multimodal fusion almost immediately after unfreezing4. Furthermore, the loss curve asymptotes smoothly towards the final epochs without significant fluctuations, which validates that our differential learning rates, specifically $\eta_{proj} = 2 \times 10^{-4}$ and $\eta_{lora} = 1 \times 10^{-5}$, successfully balanced perception and reasoning while preventing catastrophic forgetting5. This stable convergence behavior also demonstrates the efficacy of our dynamic DDP re-initialization and numerical stabilization protocols, which ensured a seamless transition between the two distinct optimization phases without encountering gradient instability or the notorious Loss NaN phenomenon.

6. Discussion

The experimental results and ablation studies validate that Lingo-Aura is a robust and effective framework for aligning low-level sensory signals with high-level cognitive reasoning. This section interprets the findings from the perspective of our working hypotheses, compares them with previous studies, and discusses their broader implications for clinical diagnostics.

6.1. Interpretation of Findings and Hypotheses

 doi:10.20944/preprints202601.0672.v1

The core hypothesis of this study was that explicit cognitive guidance and robust multimodal alignment could mitigate the "modality laziness" inherent in pre-trained large language models (LLMs). The observed 135% improvement in Pearson correlation ($Corr$) compared to text-only baselines strongly supports this hypothesis1111. In contrast to traditional Multimodal Sentiment Analysis (MSA) models that treat multimodal alignment as a simple feature concatenation task, Lingo-Aura demonstrates that injecting meta-cognitive context allows the model to "perceive" rather than just "classify" emotional states. The qualitative leap in correlation suggests that the model effectively captures the intensity of emotional leakage through micro-expressions, which is often suppressed in the verbal transcripts.

6.2. Comparison with Previous Studies

When placed in the context of existing literature, Lingo-Aura bridges the gap between high-accuracy discriminative models and high-interpretability generative models. Traditional state-of-the-art models, such as UniMSE, exhibit strong numerical fitting capabilities but function as "black boxes" that lack semantic reasoning. Conversely, zero-shot general LLMs like GPT-4 often lack the domain-specific sensitivity required for clinical diagnostics, where subtle emotional intensity is more critical than simple polarity.

Lingo-Aura achieves a unique balance by leveraging GPT-4o to generate "semantic anchors" through cognitive reverse engineering. Unlike previous attempts at multimodal fine-tuning that frequently encountered numerical instability (such as Loss NaN issues), our Double-MLP architecture with Layer Normalization provides a stable mapping into the high-dimensional embedding space. This ensures that the Mistral-7B backbone can integrate non-verbal cues without compromising its pre-trained linguistic reasoning, a synergetic effect that has been difficult to achieve in earlier multimodal adapter designs.

6.3. Clinical Implications and Broad Context

The implications of these findings are particularly significant for mental health screening and pain assessment. In psychiatric consultations, patients with conditions like Post-Traumatic Stress Disorder (PTSD) or severe depression may intentionally mask their symptoms. Lingo-Aura's ability to identify discrepancies between neutral verbal statements and micro-expression-based emotional intensity offers a powerful objective tool for clinicians to detect hidden distress.

Furthermore, in medical diagnostics, this technology addresses the critical need for pain assessment in non-verbal populations, such as infants or elderly patients with advanced dementia. By aligning involuntary facial cues with clinical context, Lingo-Aura can provide evidence-based intensity predictions that assist medical professionals in generating personalized treatment plans. This moves the field of affective computing toward "Empathic AI," where models are not only capable of processing data but are also equipped with a simulated "Theory of Mind" to understand human intent.

6.4. Limitations and Trade-Offs

Despite the qualitative improvements, several limitations remain. The global mean pooling strategy, while effective as a low-pass filter for suppressing stochastic sensor noise, inevitably sacrifices local temporal dynamics. This trade-off means that extremely transient micro-expressions lasting only a few milliseconds may be averaged out, potentially losing the specific timing information that could be diagnostic.

Additionally, we observed a minor decrease in binary classification accuracy ($Acc-2$) compared to text-only models (from 81.4% to 79.9%), which is likely due to "modality noise" from background acoustic clutter or irrelevant visual textures. This highlight the ongoing challenge of maintaining linguistic robustness while increasing sensory sensitivity.

6.5. Future Research Directions

 doi:10.20944/preprints202601.0672.v1

Future work will focus on two primary trajectories to enhance the precision and clinical utility of Lingo-Aura. First, we plan to implement advanced temporal aggregation mechanisms, such as a Q-Former or Perceiver Resampler, to capture fine-grained temporal evolutions of micro-expressions and vocal tremors. Second, we aim to integrate the framework with immersive Augmented Reality (AR) technologies for real-time, context-aware patient monitoring. Combining spatial scene understanding with hand gesture recognition will allow for a more holistic assessment of patient behavior in diverse clinical environments. Exploring multi-task learning paradigms, such as emotion cause extraction, will also be a priority to enable the model to reason about the underlying triggers of specific emotional states.

7. Conclusions

This study presents Lingo-Aura, a cognitive-informed and numerically robust multimodal alignment framework designed to enhance the perception of emotional intensity in clinical diagnostics. To address the pervasive challenges of modality laziness and noise interference in multimodal fine-tuning, we have developed a comprehensive solution that integrates low-level sensory signals with high-level cognitive reasoning. The engineering innovations of this work, including the Double-MLP projector and global temporal aggregation, have successfully bridged the modality gap while ensuring numerical stability during high-dimensional feature alignment. Experimental evaluations on the CMU-MOSEI dataset demonstrate that Lingo-Aura achieves a 135% improvement in Pearson correlation compared to text-only baselines. These results confirm that our dual-stage curriculum learning strategy and GPT-4o-generated cognitive labels effectively break the model's reliance on linguistic priors, enabling a qualitative leap from simple sentiment classification to deep affective perception.

The broader significance of Lingo-Aura lies in its potential to transform the landscape of mental health screening and non-verbal medical assessment. By aligning micro-expression features with explicit cognitive context, the framework provides a powerful tool for identifying discrepancies between verbal statements and internal emotional states, which is paramount for diagnosing conditions such as PTSD or depression. Furthermore, the robustness of the system makes it a viable candidate for pain intensity detection in vulnerable populations, such as infants and patients with cognitive impairments, who cannot communicate their distress through language. In conclusion, Lingo-Aura not only establishes a high-performance baseline for generative affective computing but also offers a replicable and cognitive-deep paradigm for the parameter-efficient fine-tuning of multimodal large language models in resource-constrained clinical environments.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
MLLM	Multimodal Large Language Model
MSA	Multimodal Sentiment Analysis
PEFT	Parameter-Efficient Fine-Tuning
LoRA	Low-Rank Adaptation
MLP	Multi-Layer Perceptron
CoT	Chain-of-Thought
CDSS	Clinical Decision Support System

 doi:10.20944/preprints202601.0672.v1

References

1. Wang, Yiqi, et al. "Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning." arXiv preprint arXiv:2401.06805 (2024).
2. Yin, Shukang, et al. "A survey on multimodal large language models." National Science Review 11.12 (2024): nwae403.
3. Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.
4. Urooj, Bushra, et al. "Large language models in medical image analysis: a systematic survey and future directions." Bioengineering 12.8 (2025): 818.
5. Ong, Jasmine Chiat Ling, et al. "Large language model as clinical decision support system augments medication safety in 16 clinical specialties." Cell Reports Medicine 6.10 (2025).
6. Zhang, Deshiwei, et al. "A survey of datasets in medicine for large language models." Intelligence & Robotics 4.4 (2024): 457-478.
7. Nam, Yoojin, et al. "Multimodal large language models in medical imaging: current state and future directions." Korean Journal of Radiology 26.10 (2025): 900.
8. Chen, Xupeng, et al. "R-llava: Improving med-vqa understanding through visual region of interest." arXiv preprint arXiv:2410.20327 (2024).
9. Chen, Qiuhui, and Yi Hong. "Medblip: Bootstrapping language-image pre-training from 3d medical images and texts." Proceedings of the Asian conference on computer vision. 2024.
10. Chen, Yirong, et al. "SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations." arXiv preprint arXiv:2311.00273 (2023).
11. Yang, Qu, Mang Ye, and Bo Du. "Emollm: Multimodal emotional understanding meets large language models." arXiv preprint arXiv:2406.16442 (2024).
12. Zhao, Xiangyu, et al. "It hears, it sees too: Multi-modal LLM for depression detection by integrating visual understanding into audio language models." arXiv preprint arXiv:2511.19877 (2025).
13. Song, Yongjia. Investigation of Cognitive Mechanisms and Factors in Bilingual Lexical Entrainment. Diss. University of California, Irvine, 2025.
14. Huang, Sining, et al. "AI-Augmented Context-Aware Generative Pipelines for 3D Content." Preprints, Aug (2025).
15. Huang, Sining, et al. "Ar overlay: Training image pose estimation on curved surface in a synthetic way." arXiv preprint arXiv:2409.14577 (2024).
16. Kang, Yixiao, et al. "6: Simultaneous tracking, tagging and mapping for augmented reality." SID Symposium Digest of Technical Papers. Vol. 52. 2021.
17. Kang, Yixiao, et al. "Tie memories to e-souvenirs: Hybrid tangible ar souvenirs in the museum." Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. 2022.
18. Huang, Sining, et al. "Immersive Augmented Reality Music Interaction through Spatial Scene Understanding and Hand Gesture Recognition." Preprints, August (2025).
19. Li, Jia, et al. "Large language models-powered clinical decision support: enhancing or replacing human expertise?." Intelligent Medicine 5.1 (2025): 1-4.
20. Zhang, Yiqun, et al. "Affective computing in the era of large language models: A survey from the nlp perspective." arXiv preprint arXiv:2408.04638 (2024).
21. Achiam, Josh, et al. "Gpt-4 technical report." arXiv preprint arXiv:2303.08774 (2023).
22. Bai, Jinze, et al. "Qwen technical report." arXiv preprint arXiv:2309.16609 (2023).
23. Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).
24. Du, Chenzhuang, et al. "Modality Laziness: Everybody's Business is Nobody's Business." (2021).
25. Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." arXiv preprint arXiv:2402.03300 (2024).
26. Zadeh, AmirAli Bagher, et al. "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.

 doi:10.20944/preprints202601.0672.v1

27. Hu, Guimin, et al. "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition." arXiv preprint arXiv:2211.11256 (2022).
28. Li, Xiang, et al. "Hybrid CNN-Mamba Enhancement Network for Robust Multimodal Sentiment Analysis." arXiv preprint arXiv:2507.23444 (2025).
29. Merians, Addie N., et al. "Post-traumatic stress disorder." Medical Clinics 107.1 (2023): 85-99.
30. Wu, Yuqi, et al. "Systematic review of machine learning in PTSD studies for automated diagnosis evaluation." npj Mental Health Research 2.1 (2023): 16.
31. Tsai, Yao-Hung Hubert, et al. "Multimodal transformer for unaligned multimodal language sequences." Proceedings of the conference. Association for computational linguistics. Meeting. Vol. 2019. 2019.
32. Liu, Zhun, et al. "Efficient low-rank multimodal fusion with modality-specific factors." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.
33. Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." International conference on machine learning. PMLR, 2023.
34. Li, Zichao. "Mcl for mllms: Benchmarking forgetting in task-incremental multimodal learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025.